COMPUTATIONAL MOLECULAR BIOLOGY
HOMEWORK 1
SHUBHRA GUPTA

In the Roche Genetics CD, there is **Introduction to Genetics** has **five** parts:
1. Genes
2. Identifying genes and their functions
3. Mutations
4. Mendelian inheritance and
5. Recombination

Genes is consists of **four** parts:

- The structure of DNA
- The concept of a gene
- Transcription and translation and
- Gene expression

Identifying genes and their functions has also **four** parts:

- Identifying genes and their functions
- Gene cloning
- Expression Studies – Pharmacogenomics and toxicogenomics and
- Model organism for genetic studies

I have divided this paper in three parts. First section contain brief history of genetic, second section contain genes and third one has identifying genes and their functions.

Brief history of Genetics:

In 1859 Darwin publishes "**The origin of species**". In 1865 Mendel publishes his work on the nature of inheritance. In 1867 Miescher discovers DNA, calling it "**nuclein**". In 1882 Fleming discovers chromosomes. In 1908 Hardy and Weinberg publish the concept forming the basis of population genetics. In 1908 – 1909 Garrod sets forth the concept of inborn errors of metabolism. In 1915 Thomas Hunt Morgan establishes the concept of linkage. In 1918 Fischer describes genetic basis to quantitative traits and introduces the concept of "**variance**". In 1926 Vogt introduces the concepts of penetrance and expressivity as aspects of variance. In 1927 Mutation was introduced by X-irradiation. In 1937 Haldance and Bell quantify linkage between 2 traits, one was color blindness and other was hemophilia. In 1940's Pauling puts forth the concept of molecular disease. In 1944 Avery, McLeod and McCarty show that genetic information is in the DNA. In 1952 Glucose-6-phosphatase deficiency found as cause of glycogen stroage disease type1: first inborn error of metabolism, in which a specific enzyme deficiency was shown. In 1953 Watson and Crick model "**the structure of DNA**" came and R Franklin carries out X-ray crystallography. In 1961 Brenner and Jacob identify the role of mRNA. In 1966 Nirenberg, Kohana and Matthei begin solving the 3-letter code of DNA - genetic code. In 1970 Arber and Smith discover restriction enzymes that cut DNA at specific sequences.

In 1977 First human gene to be cloned is chorionic somatomammotropin and also Sanger, Maxam and Gilbert publish DNA sequencing methods. In 1980 Botstein, White, Skolnick and Davis describe restriction fragment length polymorphisms (RFLPs) as linkage markers in family studies. In 1985 – 1986 "**The Human Genome Project**" was proposed and in year1986 Mullis invents the concept of PCR. In 1990 there was a official start of the "**Human Genome Project**". In 1995 first genome of H. influenza was fully sequenced. First eukaryote yeast genome was fully sequenced in year 1996. 1997 was the year of first successful attempt of animal cloning Dolly from an adult cell. In 2000 "**Human Genome Project**" completes draft sequencing the human genome.

Genetic research has taken place largely over the last 100 years. This research has importance in advance medical science. In genetic research biologists are doing research on various topics like how genes coordinates, what are their functions, what is genome, etc.

Genes:

Inside of a human cell is a nucleus. It contains chromosome that is made of **deoxyribonucleic acid** (DNA) and proteins. DNA is genetic material that is inherited. This genetic material contains the information needed by living cell, specified by structure, function, activity and interaction with other cells and environment. Genetic material contained within the nucleus (the nuclear genome), some DNA is contained in the mitochondria. The human mitochondrial genome contains 16,569 base pair (bp) of DNA including 37 genes, which encode some of the protein and RNA molecules involved in mitochondrial function. Human cell may contain thousands of mitochondria and so thousands of copies of the mitochondrial genome.

Two major kinds of nucleic acids are found in living things: **Deoxyribonucleic acid** (DNA) and **Ribonucleic acid** (RNA). Nucleic acids can be broken into monomers. These are called **Nucleotides**. **Nucleotide consists** of three parts:

1. A five carbon sugar (pentose). Two kinds of carbon sugar are found:

Ribose has hydroxyl group on its 2'. Ribose – containing nucleotides (ribonucleotides) are the monomers of RNA. Deoxyribose has a hydrogen atom on its 2'. Deoxyribose – containing nucleotides (deoxyribonucleotides) are the monomers of DNA.

2. One, two or three phosphate groups. These are attached to the 5' carbon atom of the pentose and

3. A nitrogenous base

The base is attached to the 1' carbon atom of the pentose. The nitrogenous bases of nucleotides belong to families known as **Purines** and **Pyrimidines**. In DNA, Purines are **Adenine** (A) and **Guanine** (G). Pyrimidines are **Cytosine**(C) and **Thymine** (T).

The base pairs are attached to a sugar phosphate backbone to from one of the two strands of DNA molecule. The two strands are bonded together by the base pairs.

T and G only bonds C, this results in two complementary strands. Each of strands has twisted structure. When bonded they may form a double helix. The direction of each strand in a DNA molecule is defined as 5' for the beginning and 3' for the end.
The 5' and 3' refer to the position of the nucleotide bases relative to the sugar molecule in the DNA backbone. The two complementary strands in a double helix are oriented in opposite directions. One of the most fundamental principles in DNA technology is the **Hybridization** (base pairing of two single strands of DNA or RNA or of single strands of DNA with single strands of RNA) between complementary DNA sequences to form a double stranded molecule.
The hybridization product detection and selection methods vary depending on the method, but the fundamental principle of identification is the hybridization to the best matching complementary sequence. The degree of complementarily of two sequences affects the stability of the hybridized double strand molecule. Two perfectly matched complementary sequences form a more stable double strand molecule than two sequences that are not fully complementary to each other.
The order of nucleotide bases along a DNA strand is known as the sequence. The genetic information is encoded in the precise order of the base pair. When cell divides the entire DNA is copied. For this the two DNA strands separated and new complementary strands are generated. This enable to genetic information contains with in the DNA sequence to be reproduced and transmitted to next generation. This is called **DNA replication** each strand of DNA acts as a template for the synthesis of a complementary strand.

An organism's total DNA content is known as its **Genome**. All of our cells except **Sperm** and **Eggs** cells and red blood cells carry two genome copies are called **Diploid** or 2n number of chromosomes. The diploid genome is contained with in the pairs of chromosomes. Example, the little fly Drosophila melanogaster has 8, the onion 16 and humans 46. All these numbers are even that why it is 2n. The human genome consists of 22 autosomal chromosomes that are the same in males and female. They are numbered according to their size. We also carry sex chromosomes. One of each pair inherited from parents one from father and one from mother. Males have one X and one Y chromosome. Females have two X chromosome. Overall we carry 23 pairs of chromosomes or 46 chromosomes. Each chromosome is composed of center called **Centromere**, and the two ends called **Telomere**. Centomere separates the short arm p and long arm q of the chromosome.

The reproductive germ cells or **Gametes** (that is sperm and egg cell) contain a single copy of the genome, and this is known as **Haploid** just one-half the diploid or n numbers. Sperm are tiny and the male gametes. Eggs the female gametes are larger and nonmotile. Because sperm and eggs are so dissimilar in appearance they are called **Heterogametes.** The mitochondrial genome is almost exclusively maternally inherited.

During **Zygote** formation a sperm cell contributes its nuclear genome but not its mitochondrial genome. Mitochondrial DNA is contributed to the zygote and thus the offspring by the egg only.

In each organism there is a pair of factors which control the appearance of a given characteristic. These factors called **Genes**. Genes are sequences of base pairs that encode information for proteins. They can range in size from less than 100 bp to million bp. The human gene has complex internal structure made of a portion that contains actual information for protein. These are called **Exons**. These coding portions of gene fragment are split by two stretches of noncoding DNA that contain no information. Such noncoding regions are called intervening sequences or **Introns**. In addition to exons and introns genes also contain regions that are important for regulating how actively protein is to be synthesis. Among such regulatory element our **promoters** they usually set at 5' end of gene.

All eukaryotes (i.e. organism made of nucleated cells, as opposed to bacteria which do not have a nucleus are called prokaryotes) have genes which are segmented into exons and introns. Most human genes are split into exons and introns except for mitochondrial genes and a few nuclear genes. During gene expression both exons and introns are transcribed to form an initial transcription product or pre-mRNA.
The information contain with in genes is used to make proteins by two stage methods:

1. **Transcription** and
2. **Translation**.

taken together are called **Gene expressions**.
In biological systems, the genetic information in DNA is copied into a related molecule called **ribonucleic acid** or RNA, which is very similar to DNA except for chemical modification of the sugar backbone. Also instead of Thymine (T) RNA contain **Uracil** (U), which is compare with Adenine (A).
The copying process that makes RNA is called **transcription**. Transcription begins at the start of the gene 5' (the promoter region) and continues until the end of the gene 3'. These RNA molecules are usually single-stranded and can be translated into amino acids by combining every 3 nucleotide bases (codon) along the sequence into an amino acid - which forms the building block for proteins. An RNA molecule is classified both by its cellular location and by its function. According to cellular location and function RNA divides in three major forms:

- **Messenger RNA** (mRNA) carries genetic information from DNA to ribosome.
- **Ribosomal RNA** (rRNA) 75% of the cellular component of RNA is rRNA.
- **Transfer RNA** (tRNA) carries the amino acid residue that is added to growing peptide chains (linkage between amino acid) during protein synthesis.

RNA is synthesis by unwinding the DNA double helix separating the two DNA strands and using one as a template. This is accomplice by an enzyme known as **RNA polymerase** that bind to the promoter region at the start of the gene and copies or transcribes the genes into a complementary RNA molecule.

The initial transcription product or pre-mRNA molecule is processed to form the mature mRNA. The single stranded RNA molecule is processed. The noncoding introns sequences are removed by the process called **splicing** leading the coding exons sequences join together. The splicing reaction is mediated by the spliceosome, which consists of RNA and proteins. The ends of the RNA also modified to enhance its ability. A special nucleotide is added to the 5' end of the transcript in a process known as "**capping**". Polyadenylation results in addition of a polyA tail to the transcript. The promoter region is located upstream of the coding DNA and consists of several short sequences, which are consensus binding sites for a number of proteins called transcription factors. Exons are defined as the sequences that are represented in the mature mRNA. They may or may not code for a protein: exons located at the 3' or 5' end of the mRNA may not be translated into proteins.

Translation is the process by which the genetic information contains within the sequence of nucleotide in mRNA molecule is used to synthesis protein molecule. **Proteins** are made of amino acids. There are 20 different amino acids used to build protein in humans. Each of them is code for one or more set of three nucleotides in the mRNA, so called triplet or **codons**. Translation of mRNA occurs on ribosomes. The process takes place in three phases:

1. Initiation.
2. Elongation and
3. Termination

The combination of nucleotides they build the different codons represent the genetic code. Three nucleotides (a codon) specify an **amino acid**. Since there are four nucleotides there are 4*4*4= 64 possible codons. Since there are only 20 amino acids several codons can specify the same amino acid. Therefore the genetic code is said to be **degenerate**. There are also **start codon** (Methionine, AUG), which indicate the beginning of the coding region, and three **stop codons** (UAA, UGA, UAG), which indicate the end of the coding region. All cells in the body carry the full set of genetic information, but express one point at time only 20% of the genes. Different proteins are expressed in different cells according to the function of the cell.

Identifying genes and their functions:

Identifying and characterizing a gene is to isolate from rest of the genome and produce large enough quantity of it to allow its investigation, this process is called **cloning** a gene. There are two basic approaches to DNA cloning: cell based DNA cloning or cell free DNA cloning (PCR). Basic example is to clone DNA fragment using bacteria. In this case the DNA fragment containing the gene of interest was isolated from the entire genome by using specific enzymes called **restriction enzymes** is inserted into a vector using DNA ligase (enzyme) and the recombinant product is introduced into bacteria, which make new copies with every cell division. These are specialized enzymes they allowed the side specific cutting of a DNA at a particular sequence. Typically restriction enzymes recognize specific DNA sequences of 4, 6 or 8 bases in length.

The DNA fragment is then inserted into so called vector (like mini chromosome) and introduced into bacteria. This process is called transformation. As the bacteria multiply the vector with the DNA is also multiplied to produce many copies of this DNA. Then can be isolated in large quantity from the bacteria. Cloning vectors are DNA fragments that are able to replicate within a cell and allow the addition of exogenous DNA. Most cloning vectors are derived from plasmids (mini chromosomes), viruses, phages or chromosomes. Vectors are classified according to properties such as the type of host cell they can replicate (e.g. bacteria, mammalian cell) or the size of the exogenous DNA.
A gene extracted from DNA form the nucleus called gnomic DNA contains all the known non-coding intronic sequences. Therefore it is very large and difficult to analyze for function. And alternate is start with mRNA, which contains only the exons or coding sequences, and therefore it is simple to study. Before cloning mRNA is converted into DNA which is called **complementary DNA** or cDNA. cDNA are DNA molecules that are complementary to the mRNA sequences in the sample. cDNA is synthesized by the enzyme reverse transcriptase (RT) that uses the mRNA as template.

In order to characterize the function of gene it is important to know its sequence and comparison with other sequences in database. Identify where and under what condition gene can be expressed. And what function is known it has in other organism.
Gene expression allows you to understand how a gene is regulated in tissue or cell type. Technically it can be measured by the level of mRNA produce from particular gene in a particular tissue. One of the most common techniques to analyze the mRNA level of single gene is called **Northern Blot**. In most precise way to quantify gene expression is **quantitative reverse transcriptase** PCR-RT-PCR. In addition DNA array allow the studies of genes expression for 10's of 1000's of different genes from one tissue sample in one experiment.

DNA Microarray:

DNA microarrays consist of thousands of DNA probes (corresponding to different genes) arranged as an array. Each probe is complementary to a different mRNA (or cDNA). The mRNA, which is isolated from a tissue or cell type, is converted to fluorescently labeled cDNA and used to hybridize the array. All expressed genes in the sample will bind to one probe of the array and generate a fluorescent signal.
This method provides advantage (over other mRNA profiling method) that it can interrogate the level of transcription of several thousands, of different genes from one sample in one experiment. The array reverses the principle of the Northern Blot: the probe are fixed and not labeled, while all RNA's in the sample are labeled and the sample is in the solution with which the microarray is incubated.

There are different types of DNA arrays designed for RNA profiling. These can differ by the type of probe immobilized on the array (e.g. plasmid DNA or synthetic oligonucleotides) the number of genes represented on the array and the density (probes per square centimeter of array) of the array. DNA chip microarray are density microarrays, in which the probes consist of synthetic DNA fragment (oligonucleotides (are short sequences of single stranded DNA or RNA) which are synthesized directly onto the chip much as it is done with microprocessor chips. Today a DNA microarray the size of a fingernail can interrogate far more than 10,000 different transcripts. The chip carries between 30 and 40 different probes per transcript. Half of the probes interrogate each gene are designed to perfectly match 20 nucleotide stretches of the gene, while half of them contain a mismatch (a faulty nucleotide) as a control to test for specificity of the hybridization signal.

Pharmacogenomic of Melanoma Tumor:

Skin melanoma tumors can be successfully treated with alpha interferon. However, a large percentage of these melanomas show no response to the treatment and these patients only experience the side effect of the drug. The mechanism for the interferon resistance and final a diagnostic that would allow prediction regarding the response to the drug, an mRNA profiling study was done on melanoma tumor cell lines derived from patients who did and did not respond to the treatment. When exposed to ∝-interferon treatment the cell lines from responders showed reproducibly an up regulation of certain genes that was absent in the cell lines form non-responders, and vice-versa.

Toxicogenomics:

Toxicogenomics is the study of the influence of toxic substances on gene expression in one or several tissues. Toxicogenomics provides new ways of examining the potential toxicity of a drug, identifying which genes are increased or decreased in their expression in particular organ due to influence of a drug allow us to understand better whether drug may have advert effects and why.

Model organisms are indispense tool to study the function of a gene. These model organism are range from Bacteria, yeast, worms(C. elegans), insect cells, frog eggs (Xenopus oocytes), files (Drosophila melanogaster), zebra fish, mice and the most difficult is mammalian (human). In general the more complex the organism is more difficult it is due to genetic model.

Way to test function of a gene is to inactivate it expression in mice, using the so called **Knock-out technique** or introduce a new gene by generating a transgenic animal. Knock-out genes are inactivated by a directed mutation. Knock-in a gene is introduced in a specific locus to replace another gene. Transgenes a gene is introduced somewhere, at random in the genome.

Public Databases for DNA Sequences:

There are varieties of public databases from which DNA sequences can be retrieved. The **GenBank** database contains all public DNA and protein sequence data from humans and a growing number of other organisms.

Human Genome Project (HGP):

The Human Genome Project is an international research effort to map the human genome and the genomes of other organisms. It officially began in 1990 with planning and funding through the US Department of Energy and the National Institutes of Health. The aim was to complete physical and genetic maps of the entire human genome.

DNA Sequencing:

DNA sequencing is the process designed to precisely determine the sequence of bases in the DNA. Currently DNA sequencing involves enzymatically copying the DNA in the presence of compounds that terminate this copying process in a base specific manner. This results in a mixture of DNA copies that differ in size by one base. Different technologies are used to resolve the mixture, and detect the different fragments.

Polymerase Chain Reaction (PCR):

PCR allows the selective amplification of DNA sequence. DNA's to be amplified is 500 bp long sequences from the human insulin gene. The source of DNA for amplification can be any tissue. Genomic DNA was isolated from the blood sample. Only tiny amount of DNA is necessary to obtain a PCR product. In fact a blood drop or less would be enough material to start with. The DNA sequence flanking the target sequence, in this case the region of insulin gene need to be know in advance. So set of complementary DNA primer scan be designed. **Primers** are short synthetic DNA sequences of about 20 bases called **oligonucleotide**. They can specially hybridize to unique complementary DNA sequences. Genomic DNA, primers, the template, the starter, dioxynucleotices, the building blocks and very special DNA polymerase they must be resistance to heat, the motor of the reaction are all mixed together in one reaction tube. The reaction takes place in thermocycler and apparatus that allow one to precisely heat and cooled the reaction. The DNA is heated to almost bowling temperature which separate the two strands of double helix a process called heat **denaturalization.** By cooling the mixture the primer are then allowed to specially bind to the complementary sequence in the genomic DNA. Once the primer binds the DNA polymerase using them as start sides to generate a copy of each strands of a insulin gene fragment does building the two new double stranded molecule. These polymerase product are then again heat natured and separated strand cooled to allow a new set of primers to bind the DNA polymerase now generates two more copies of each strand. After 30 such cycle of denaturalization and polymerase reaction about one billion new copies of genes has generated. This is enough material to do further analysis on the gene.

To understand certain biological processes it is useful to study the differences in gene expression, which occur during such processes. For example it may be of interest to know which genes are induced or repressed in a particular organ, e.g. the liver, after an individual takes a particular drug or which genes are expressed in a tumor but not in the surrounding normal tissue.